Directions and Issues for High Data Rate Wide Area Network Environments

or

The China Clipper Project: A Data Intensive Grid

Support for Dynamically Configured, Adaptive, Distributed, High-Performance Data and Computing Environments

Gary Hoo, William Johnston, Brian Tierney, and Craig Tull, LBNL; Jim Leighton LBNL/ESNet;
Andrew Hanushevsky, Dave Millsom, and Richard Mount SLAC;
Ian Foster, ANL/U. Chicago; Alain Roy, U. Chicago

1935: Pan American's "China Clipper"
-- a Martin M-130 --
inaugurates Pan Pacific passenger service, flying from San Francisco to the Orient.

Like these "flying boats" at the beginning of large-scale airline service, the China Clipper Project anticipates future data-intensive environments in both flexibility and performance.

wejohnston@lbl.gov

This work is supported by the Director, Office of Energy Research, Office of Computation and Technology Research, Mathematical, Information, and Computational Sciences Division, of the U. S. Department of Energy under Contract No. DE-AC03-76SF00098 with the University of California.

Data-Intensive Computing in Widely Distributed Environments

Motivation

- **Major scientific instrumentation systems -- synchrotron X-ray sources at LBNL, ANL, and BNL, the gigahertz NMR systems at PNNL, high energy particle accelerators at half a dozen DOE labs (SLAC, Fermi, ORNL, LANL, etc.) -- are all national user facilities: They involve scientific collaborators throughout the country, and internationally.**
- **These instruments generate massive amounts of data, at high rates (hundreds of megabits/sec, and more), and all of this data requires complex analysis. (E.g., see [3], [5], [7] and [8].)**
- **This data must be analyzed by scientific collaborations at the DOE Labs and at hundreds of universities.**

Motivation

Increasingly science is done in multi-instrument, multi-institutional, multi-disciplinary environments that are always distributed: The required resources are never all in the same place, and some aspects must be globalized.

- **The hundred thousand data tape archive systems needed to organize and preserve the data from such collaborations are typically maintained at a few large facilities that are commonly located at sites other than the instruments, and therefore must be accessed remotely.**
- **The intellectual and computing capacity to analyze the data is distributed among the sites participating in the scientific collaborations.**
- **Data whose life extends beyond a single experiment are maintained, annotated, and catalogued by discipline experts at their home institutions.**

Motivation

Thus, organizing, analyzing, visualizing, and moving around the nation, massive amounts of data, as well as employing large-scale computation is essential to the modern scientific process, and depends on uniformly available, distributed services to aggregate and utilize the required computing and storage components.

Furthermore, this scenario for data analysis is just the beginning.

Motivation

The next major advances in the data intensive environments, and the corresponding advances in scientific experimentation, will come in two related scenarios:

- **The first advance will come when we are able to do real-time analysis of instrument data so that scientific experiments may be validated, modified, and controlled based on "immediate" feedback to the scientists operating the instruments.**

- **The second advance will come when we can directly couple the instruments and the associated distributed, high performance data analysis environments envisioned here, to the computational simulations of the underlying physics, chemistry, materials science, etc., so that the interplay between experiment, theory, and scientific insight can happen in an integrated and interactive environment.**

Background: Example Architecture for On-line Data Sources

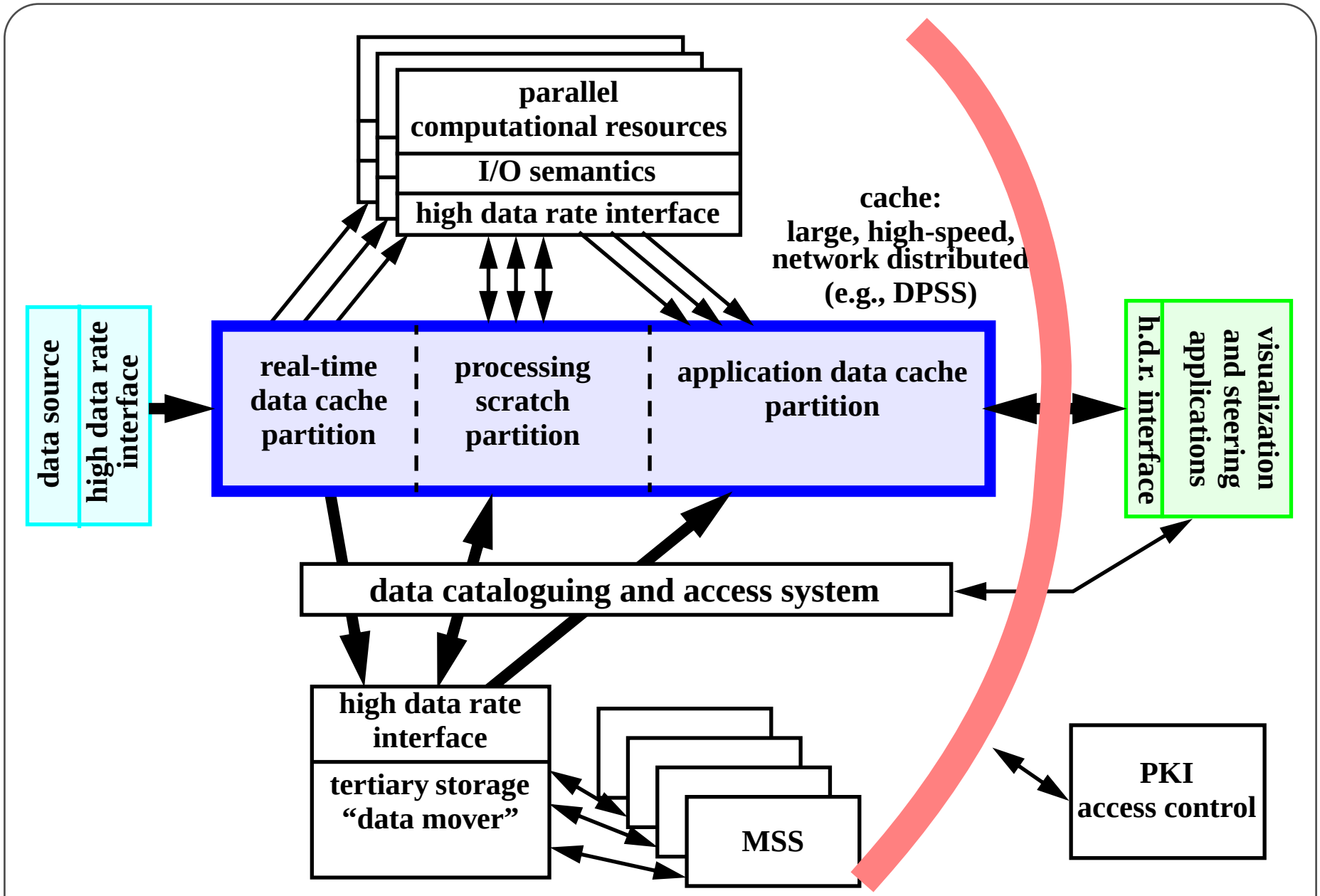
From our experience in with on-line instruments, a prototype architecture has evolved from:

- **the Kaiser experiments with network caches and on-line angiography (x-ray video) systems**
- **simulated on-line HENP detectors**
- **distributed satellite image processing in MAGIC testbed**

Key features of the architecture:

- **very high-speed cache that is distributed, scalable, and dynamically configurable**
- **common, low-level, high data rate interface that supports various application I/O semantics**
- **high-speed tertiary storage interface**
- **data cataloguing and access system**
- **distributed management of strong access control**

Example Architecture



A Prototype High Volume, High Data Rate, Data Analysis Architecture

A Data Intensive Computational *Grid*

From the application's point of view:

A layered collection of middleware services that provide to applications uniform views of distributed resource components and the mechanisms for assembling them into systems.

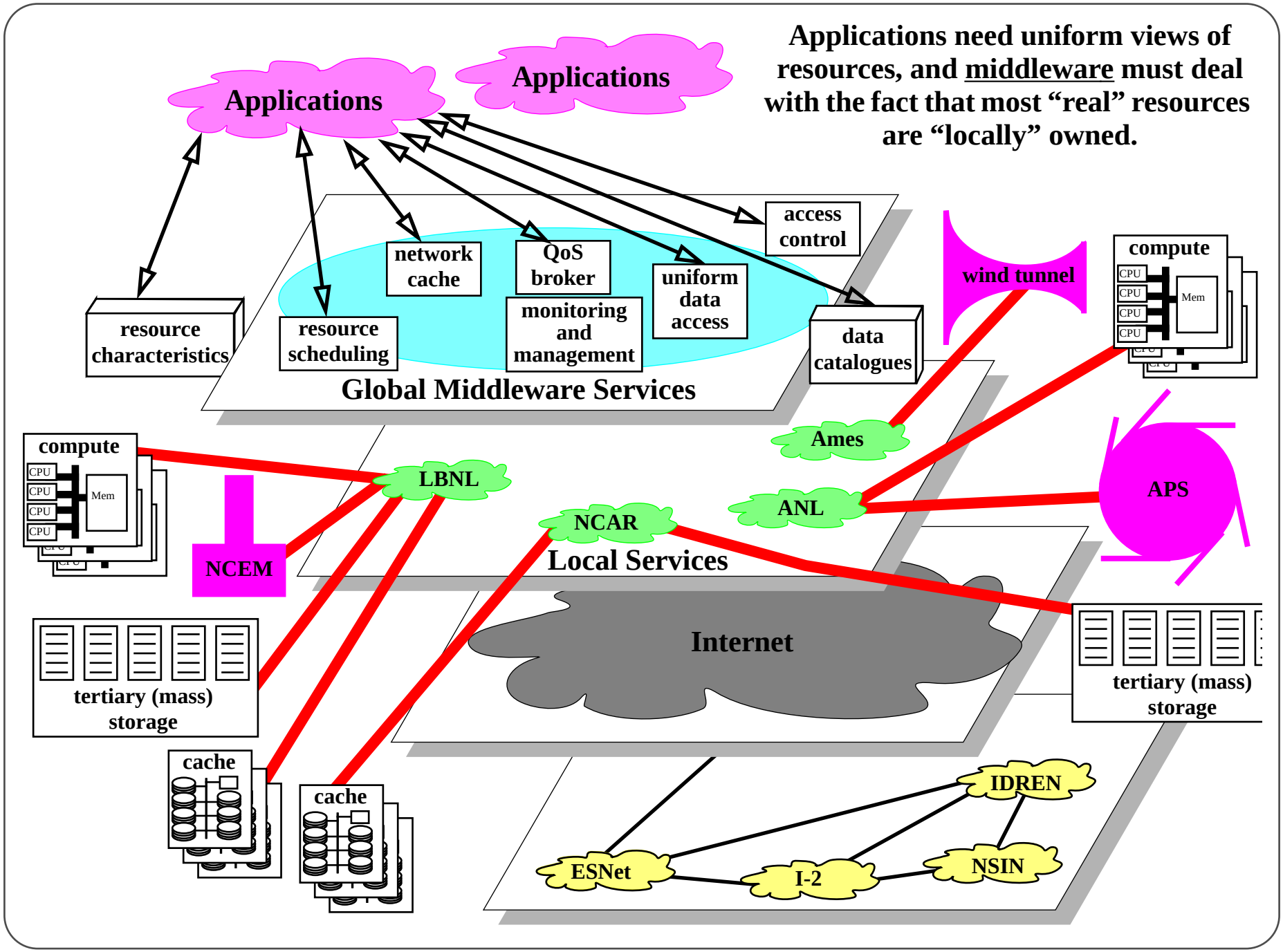
From the Grid[1] implementer's point of view:

These services extend both "up and down" through the various layers of the computing and communications infrastructure. I.e., they will "drill through" layers when necessary to achieve the required performance.

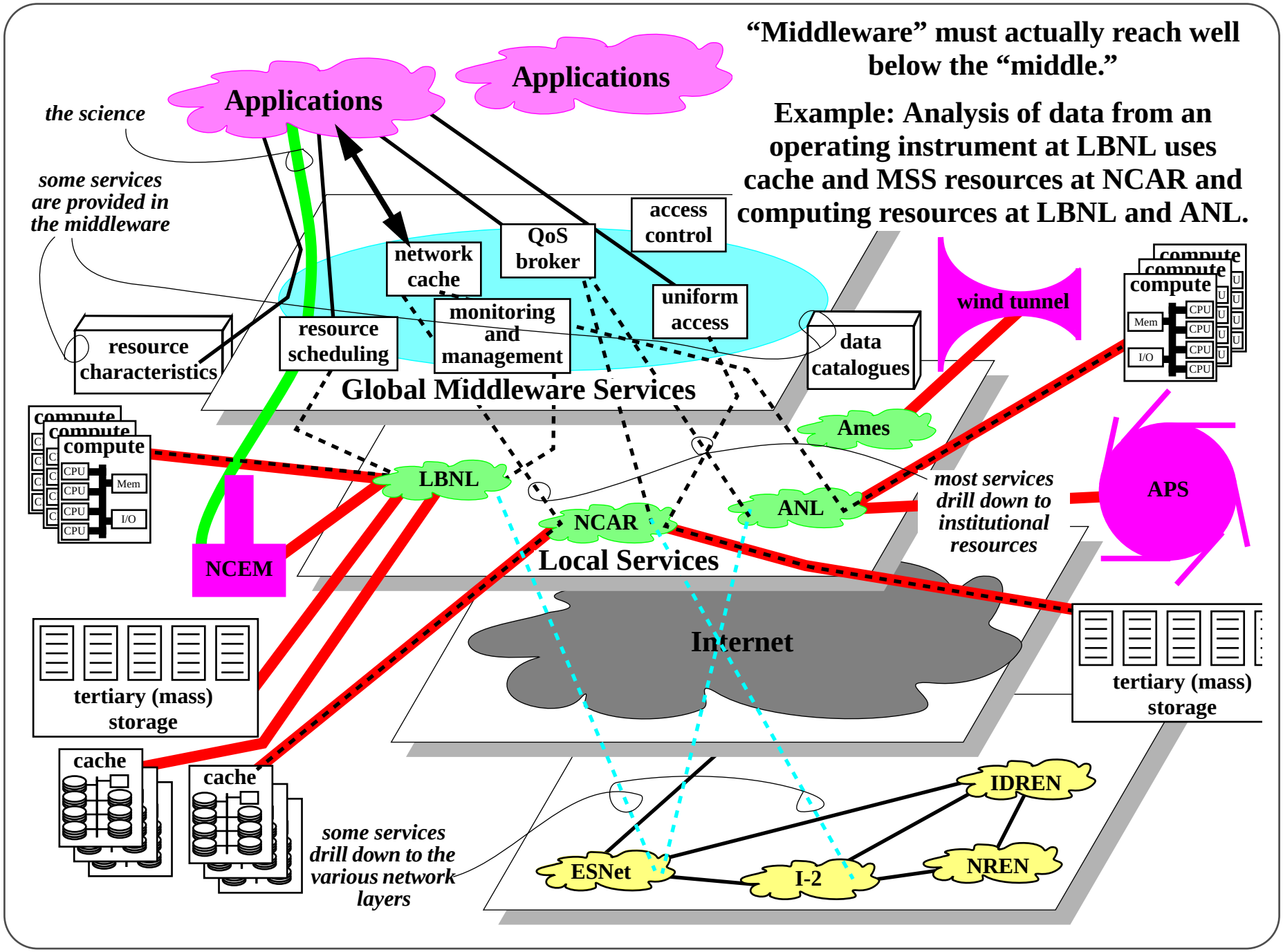
---

1. The term "Grid" is an appropriate (and increasingly used) description of generalized and widespread computing and data handling infrastructure. See, e.g., "The Grid: Blueprint for a New Computing Infrastructure." [18]

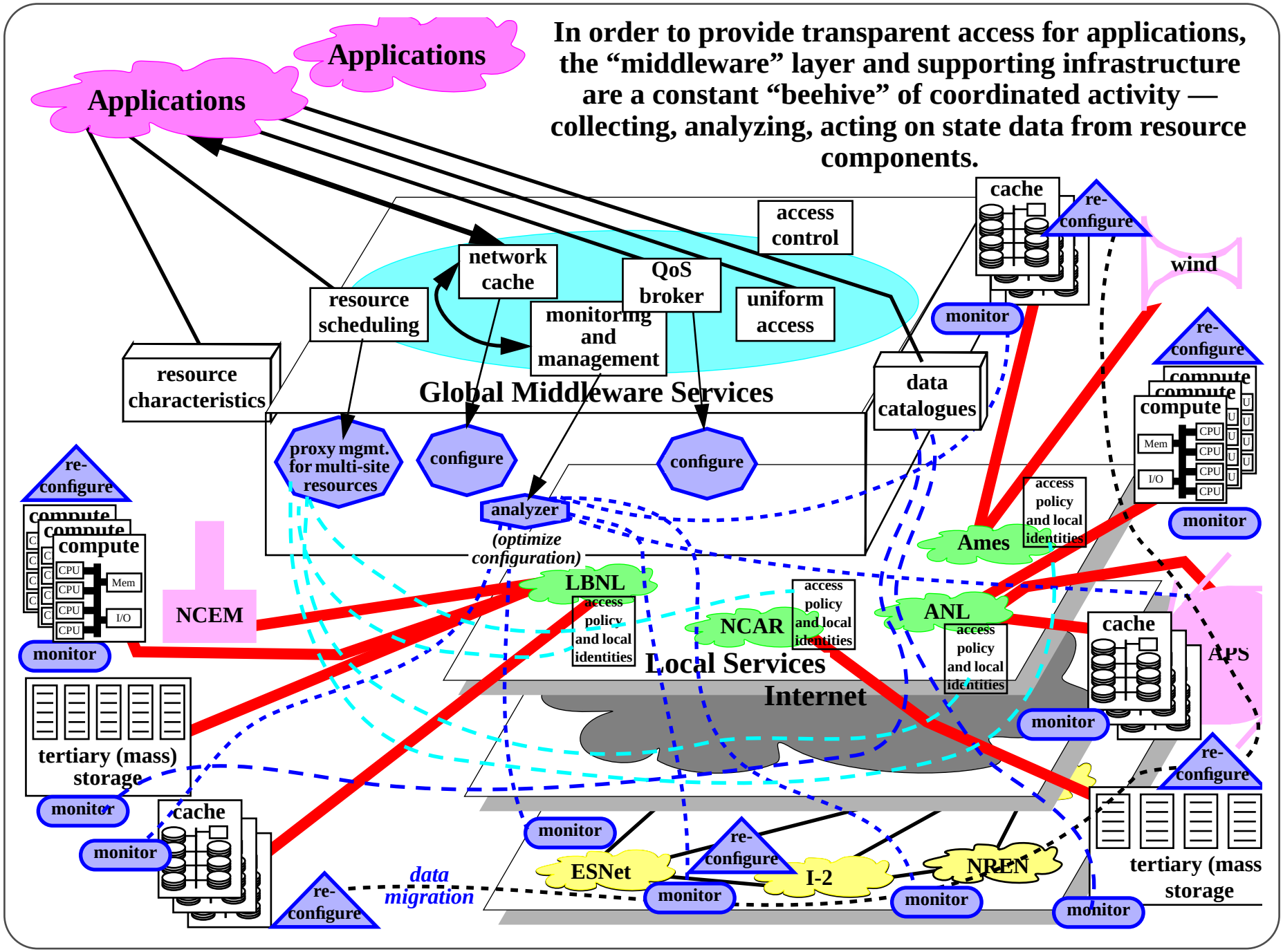
General Approach



General Approach



General Approach



General Approach

The Middleware Issue

What problem is being solved (either implicitly or explicitly) by a particular middleware model?

- **Multi-disciplinary data analysis based on catalogued datasets (e.g. MDAS)**
  - **large number of unrelated users**
  - **substantial effort in metadata, etc.**
- **Single discipline analysis of very large (and growing) well structured data sets (e.g. STAF / HENP)**
  - **e.g., data from a single class of instruments**
  - **homogeneous user community (e.g. high energy physics)**
- **Automated acquisition and cataloguing of metadata-rich data sets (e.g. WALDO)**
  - **imaging systems**
  - **mixed, but limited user community**

General Approach

- **"Second tier" communities that use storage and computing resources wherever they can get them (e.g. Globus)**
  - **everything is scattered**
  - **users = much of the scientific community that does modeling**

Systems such as SDSC's Massive Data Analysis Systems (MDAS), Globus, Legion, WALDO, and STAF are all "tuned" - to a greater or lesser extent - for a particular model of resource usage (i.e. a particular community).

Middleware also supports different programming models:

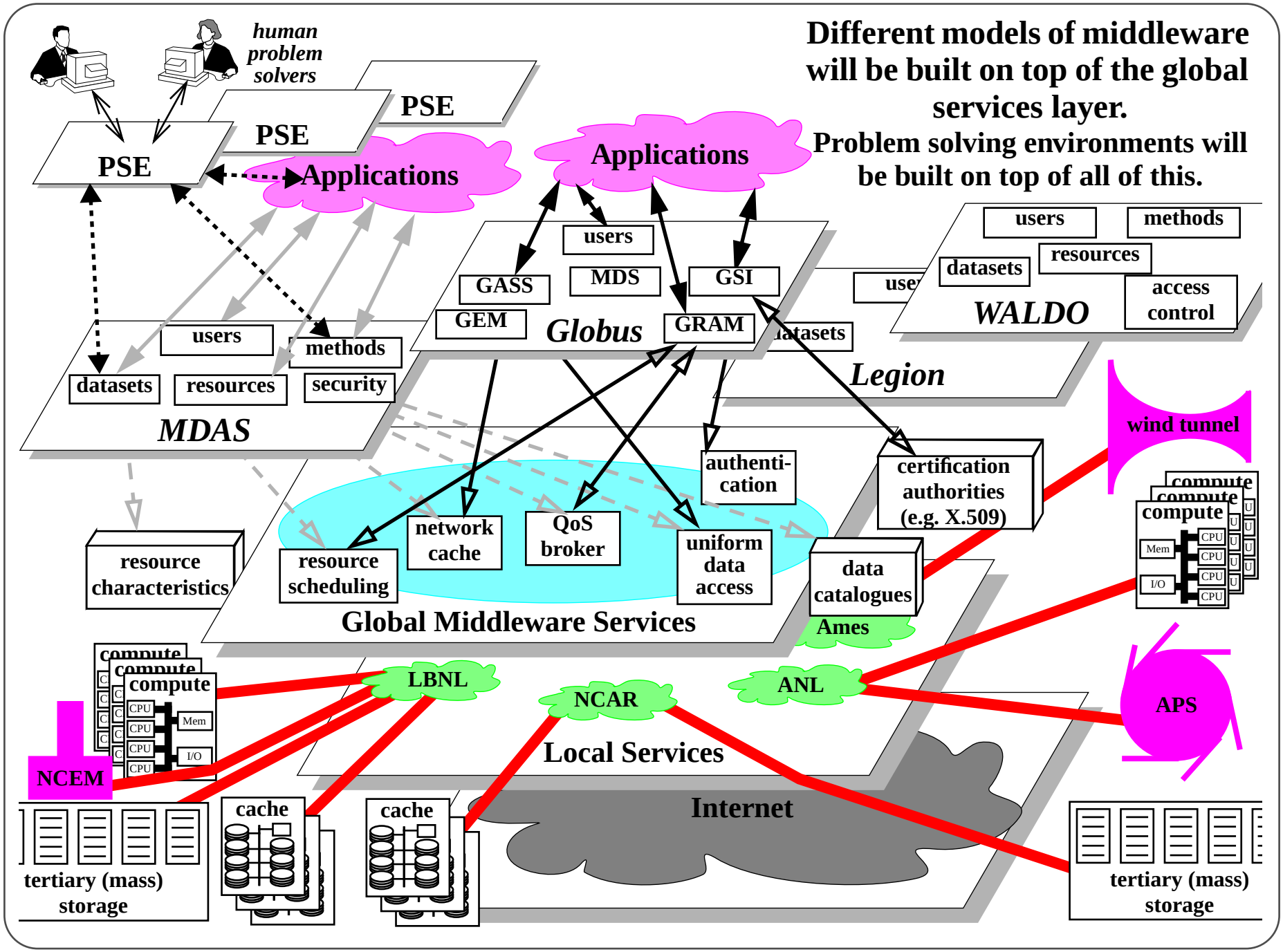
- **Globus provides distributed, shared memory via MPI**
- **PVM provides distributed, shared memory via RPCs**
- **Legion provides distributed objects**

General Approach

While there are many common elements in the middleware models, each system emphasizes different elements. Some elements may be complicated and of great interest to one community, and of little or no interest to another, who will therefore be unwilling to bear the cost or complexity. Similarly, certain models may lack elements of critical interest to a community that then sees this model as useless.

- **There will not be a single model for defining, structuring, and implementing middleware.**
- **However, there probably will be a set of infrastructure services can be used to support the construction of various middleware systems.**
- **The components of the Clipper project are probably most appropriately called infrastructure/Grid services for building middleware.**

General Approach



Where We Are Today (in a demo Grid)

The LBNL $\leftrightarrow$ NTON[2] $\leftrightarrow$ SLAC experiments:

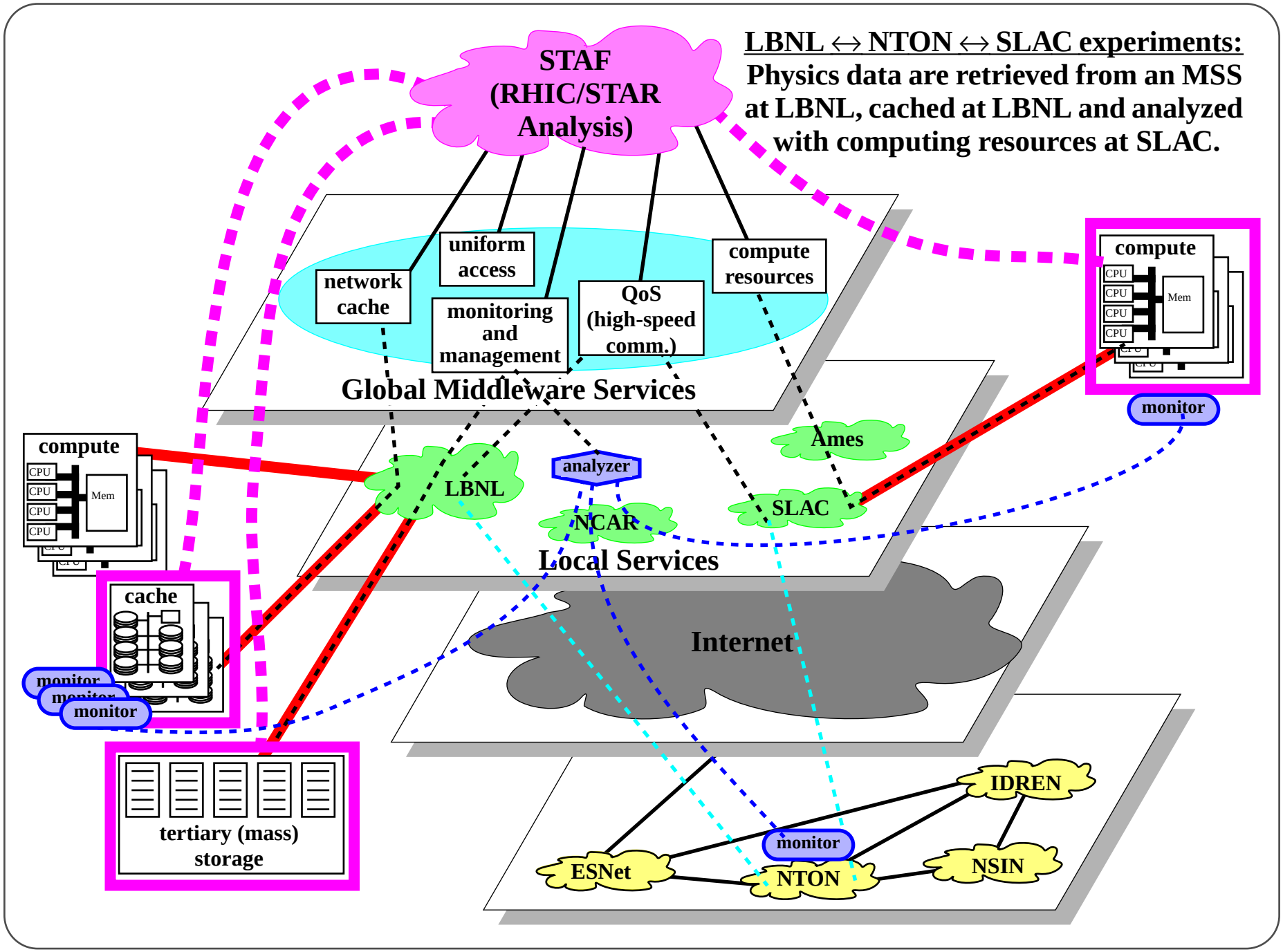
An application reading data from a remote distributed-parallel, network based cache and decoding it into application memory— at almost 60 Mbytes/sec sustained — provides a proof-of-principle.

However, while a lot of progress has been made over the past five years toward high performance wide-area data-intensive computing, what we still have is a collection of disconnected, prototype technologies.

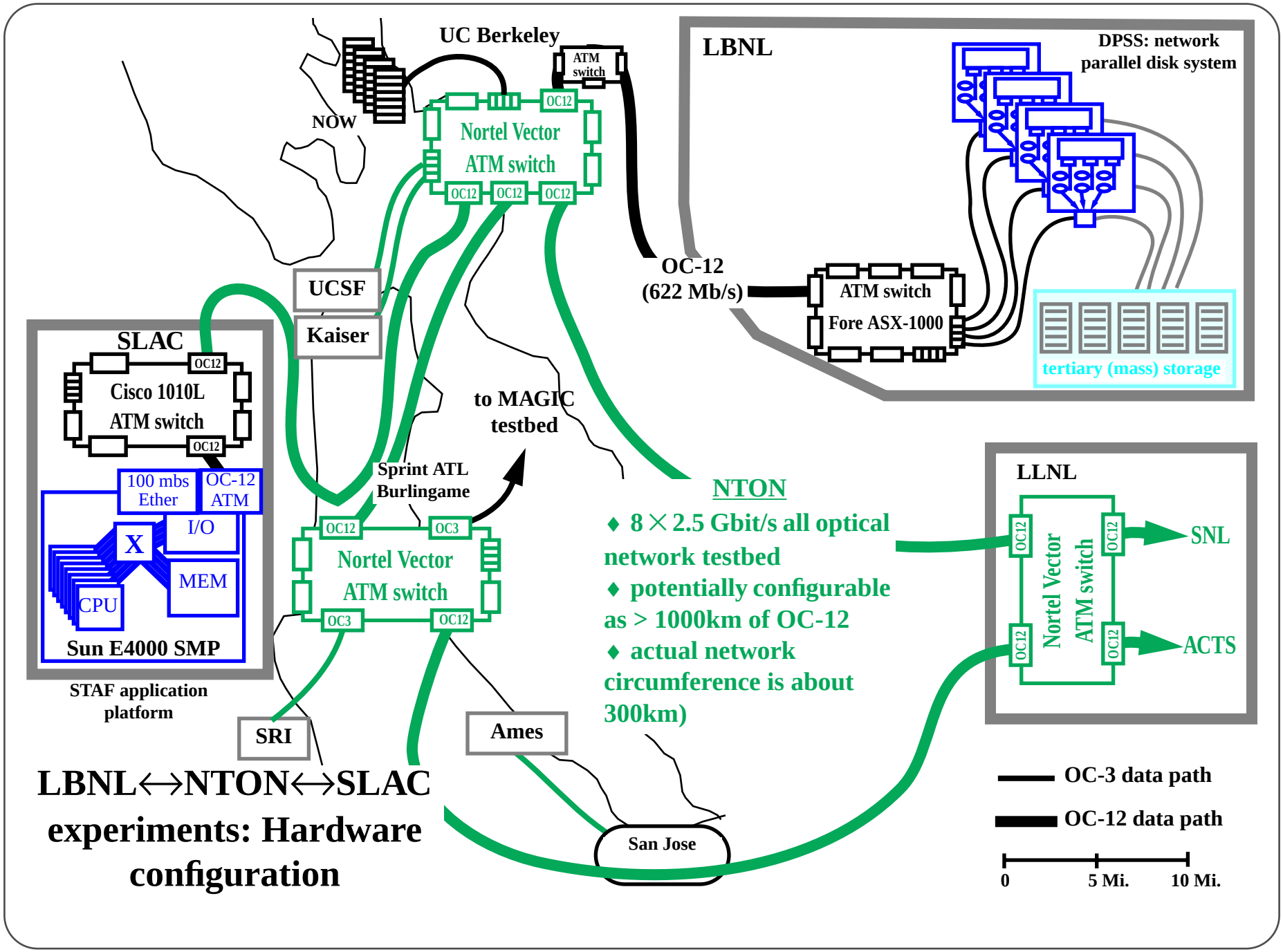
---

2. National Transparent Optical Network testbed. See [23].

Where We Are Today

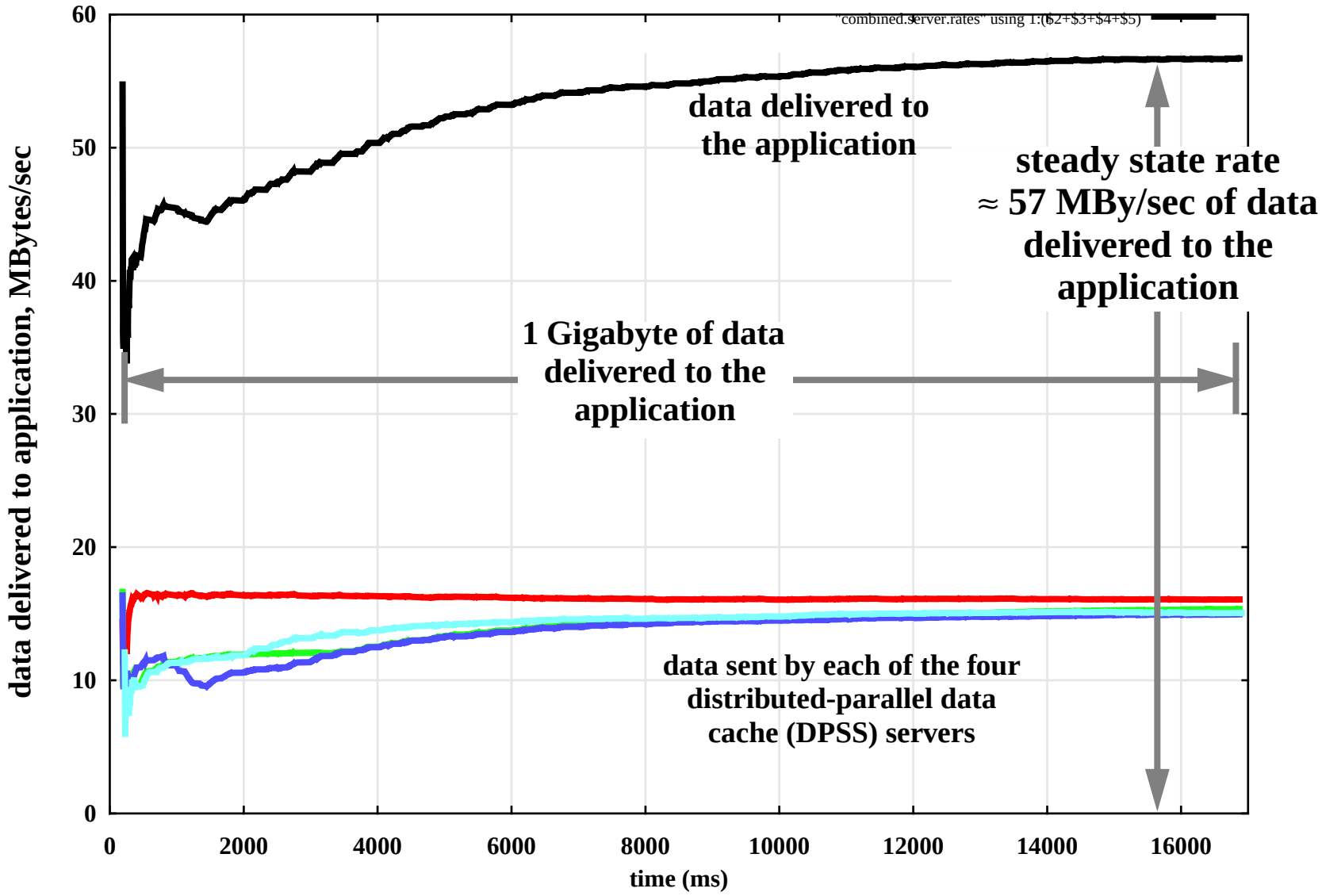


Where We Are Today



Where We Are Today

NetLogger Trace: Four DPSS server test, LBNL <-> SLAC via TCP over NTON



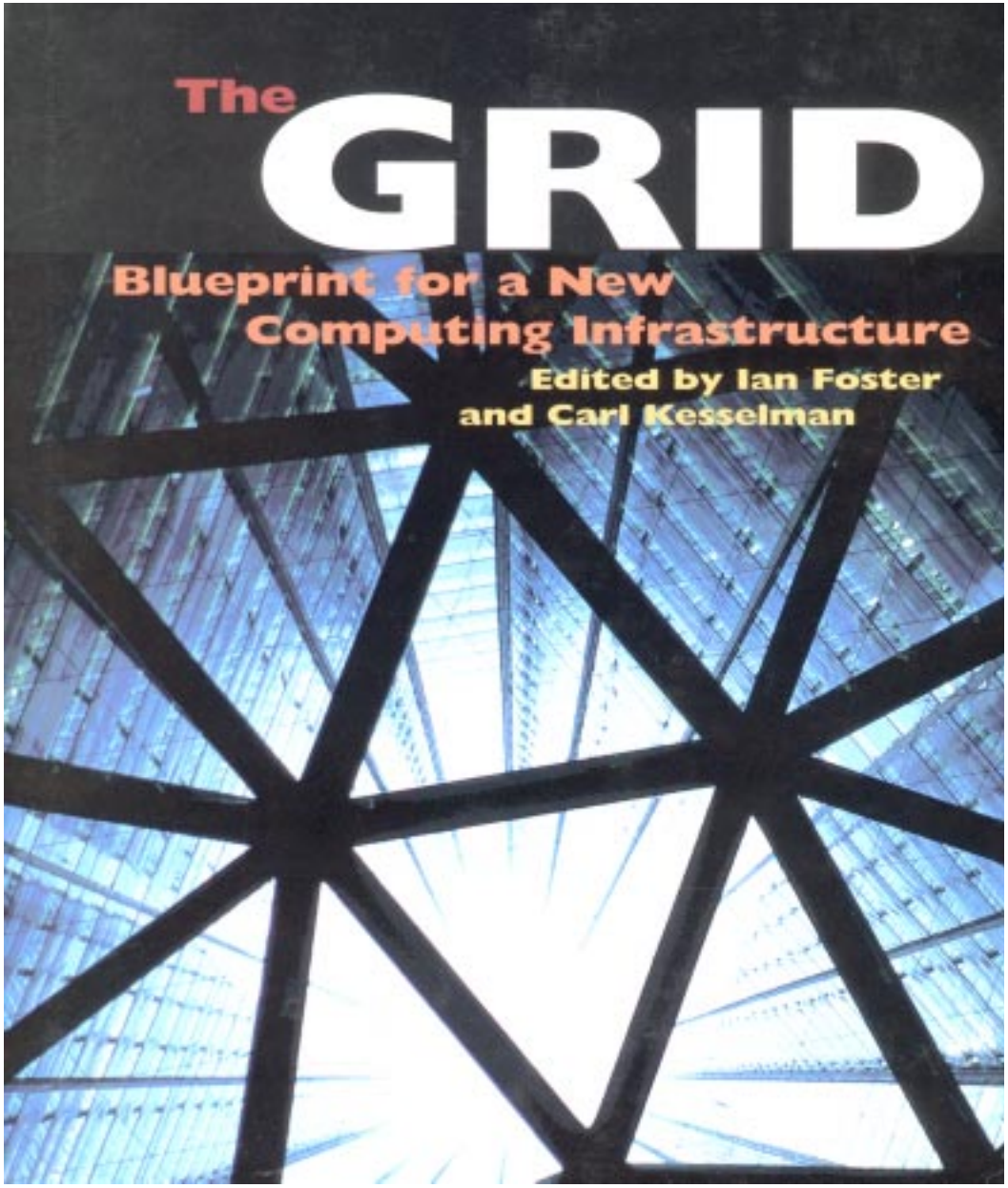
Experiment: Application reading data from four remote servers.

Building the Grid

- **What are the services that are common to all middleware systems?**
- **What are the issues for the routine creation of useful, robust, high-data rate, distributed applications?**
- **What are the states of the current testbeds?**
- **What are the prospects for building persistent grid environments?**

Where We Are Today

This book[*] [18] explores the current ideas and technology for building general, persistent, large scale, and routinely available distributed computing infrastructure.



([*]Although an author, WEJ has no financial interest in the book. All of the authors have agree to donate any royalties to a student scholarship fund for the IEEE High Performance and Distributed Computing conference.)

Where We Are Today

What are the Grid services that are common to all middleware systems?

- **Technology for coupling distributed computing systems to high data-rate on-line instruments for “rapid” data analysis and real-time control**
- **High-speed network connectivity that offers schedulable quality of service to effectively provide some level of bandwidth reservation among the elements of widely distributed systems**
- **High-performance network access to tertiary storage (“MSS”)**
- **Data management systems that provide global name spaces and federated views of multiple archival mass storage systems**

Where We Are Today

- **High-speed caches that provide applications with very high-performance access to data staged from MSS, or in various stages of analysis, regardless of the physical location of the data or grid elements accessing it**
- **Comprehensive precision monitoring to support diagnosis and infrastructure performance experiments, and to provide performance trend indicators to adaptive applications**
- **Active management mechanisms providing fault-tolerant operation and dynamic reconfiguration of widely distributed system elements**
- **Programming services that support distributed-parallel code development, debugging, and execution in the wide area**
- **Resource catalogues and global name spaces for data**

Where We Are Today

- **Job management services such as distributed scripting mechanisms, check-point and restart, automated cataloguing of execution output, execution status monitoring, etc.**

- **Security:**
  - **access control infrastructure that supports automated, policy based scheduling of the resources required by an application, and provides strong enforcement of diverse administrative use-conditions and scheduling agreements**
  - **security services to provide authenticated and/or confidential communication (e.g. Secure Sockets Layer for streams and the Generic Security Services for messages)**
  - **infrastructure protection mechanisms such as secure DNS and IPsec to authenticate packet senders (and reject unauthenticated senders), secure routing updates and router and switch maintenance ports, etc.**

Where We Are Today

What are the states of the current testbeds?

- **Gusto - A Globus Grid**

  "The Globus project is currently building two testbeds. Our main testbed is called the Globus Ubiquitous Supercomputing Testbed, or GUSTO. Gusto currently has twenty-seven participating sites and over 2.5 TFLOPS of compute power. GUSTO represents one of the largest computational environments ever constructed."

  "Our second testbed is being used to conduct experiments in network quality of service. This testbed is layered on top of CAIRN, an experimental network primarily sponsored by the Information Technology Office of DARPA."

  http://www.globus.org

Where We Are Today

- **Clipper - A data intensive grid**

  Clipper is focused on data intensive applications. The network infrastructure consists of OC-12 ATM links between LBNL, SLAC, and LLNL (via the NTON testbed [23]) and an OC-12 link between LBNL and ANL (ESNet).

  The monitoring, distributed caches, network QoS, and techniques for high-speed network access to mass storage systems (e.g. HPSS) needed for data intensive applications are being integrated with Globus.

  http://www.itg.lbl.gov/Clipper

Where We Are Today

- **Information Power Grid ("IPG" - NASA) - A generalized Globus Grid (coming soon)**

  NASA's IPG is a new project at the Ames Research Center that has the goal of making Grid architecture systems the standard, ubiquitous computing and data infrastructure for NASA.

  http://ipg.arc.nasa.gov

Where We Are Today

What are the prospects for building persistent grid environments?

As a result of annual Grid workshops where the middleware developers discuss architecture and implementation issues, IETF-like forums are being set up to standardize common sets of "global middle services).

NASA is working to put a Grid into production use in support of aerospace engineering systems.

Notes

Material that directly complements this talk — usually with design and/or implementation details — may be found in [19].

Unless otherwise noted, these paper are on the Web, and pointers may be found at http://www-itg.lbl.gov/~johnston/papers .

[1] "High-Speed, Wide Area, Data Intensive Computing: A Ten Year Retrospective,"
William E. Johnston. 7th IEEE Symposium on High Performance Distributed Computing, Chicago, Ill. July 29-31, 1998.

[2] "The NetLogger Methodology for High Performance Distributed Systems Performance Analysis,"
Brian Tierney, W. Johnston, J. Lee, G. Hoo, C. Brooks, D. Gunter. 7th IEEE Symposium on High Performance Distributed Computing, Chicago, Ill. July 29-31, 1998.

[3] "Real-Time Widely Distributed Instrumentation Systems,"
William E. Johnston. In *The Grid: Blueprint for a New Computing Infrastructure*. Edited by Ian Foster and Carl Kesselman. Morgan Kaufmann, Pubs. August 1998.

[4] "Authorization and Attribute Certificates for Widely Distributed Access Control,"
William Johnston, S. Mudumbai, and M. Thompson. IEEE 7th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises - WETICE'98, Stanford, CA. June, 1998.

[5] "Real-Time Generation and Cataloguing of Large Data-Objects in Widely Distributed Environments,"
W.Johnston, Jin G., C. Larsen, J. Lee, G. Hoo, M. Thompson, and B. Tierney (LBNL) and J. Terdiman (Kaiser Permanente Division of Research). Invited paper, International Journal of Digital Libraries - Special Issue on "Digital Libraries in Medicine". May, 1998.

[6] "Research & Development Priorities for Communications and Information Infrastructure Assurance,"
William J. Hunteman (Los Alamos National Laboratory), Sharon E. Jacobsen (Oak Ridge National Laboratory), William E. Johnston (Lawrence Berkeley National Laboratory), Douglass L. Mansur and Kathleen C. Bailey (Lawrence Livermore National Laboratory). Presidential Commission on Critical Infrastructure (PCCIP) taskforce report. June, 1997.
http://www-itg.lbl.gov/security/PCCIP/C+I_Report.html

[7] "High-Speed Distributed Data Handling for On-Line Instrumentation Systems,"
William E. Johnston, W. Greiman, G. Hoo, J. Lee, B. Tierney, C. Tull, and D. Olson. Proceedings of ACM/IEEE SC97: High Performance Networking and Computing. Nov., 1997.

[8] "High-Speed Distributed Data Handling for High-Energy and Nuclear Physics,"
William E. Johnston, W. Greiman, B. Tierney, A. Shoshani, and C. Tull. Invited lecture (W. Johnston) at the 1997 CERN School of Computing, Pruhonice (Prague), Czech Republic. 17 - 30 August 1997. Published in the "Proceedings of the 1997 CERN School of Computing." Also published in the proceedings of Computing in High Energy Physics, Berlin, Germany. April, 1997.

Where We Are Today

[9] "Performance Analysis in High-Speed Wide Area ATM Networks: Top-to-bottom end-to-end Monitoring,"
B. Tierney, W. Johnston, J. Lee, G. Hoo, IEEE Networking, May-June, 1996.

[10] "A Distributed Parallel Storage Architecture and its Potential Application Within EOSDIS,"
W. Johnston and B. Tierney (LBNL), and J. Feuquay and T. Butzer (USGS EROS Data Center). NASA Mass Storage Conference, Goddard Space Flight Center, April, 1995.

[11] Akenti
PKI, Attribute, and Use-Condition certificate based access control with distributed management of multi-party policy.
See http://www-itg.lbl.gov/security/Akenti

[12] CIF
The CIF project aims to develop a distributed computing software bus which supports software components for scientific collaboratories developed by diverse DoE and non DoE groups.The goal of the project is to reduce duplication of effort, enhance interoperability and promote interlab cooperation by developing a common communication library for all projects.
See http://www-itg.lbl.gov/CIF

[13] Clipper
The goal of the Clipper project is software systems and testbed environments that result in a collection of independent but architecturally consistent service components that will enhance the ability of applications and systems to construct and use widely distributed, high-performance data and computing infrastructure. Such middleware should support high-speed access and integrated views for multiple data archives; resource discovery and automated brokering; comprehensive real-time monitoring and performance trend analysis of the networked subsystems, including the storage, computing, and middleware components, and; flexible and distributed management of access control and policy enforcement for multi-administrative domain resources.
See http://www-itg.lbl.gov/~johnston/Clipper

[14] diffserv
"There is a clear need for relatively simple and coarse methods of providing differentiated classes of service for Internet traffic, to support various types of applications, and specific business requirements. The differentiated services approach to providing quality of service in networks employs a small, well- defined set of building blocks from which a variety of services may be built."
http://www.ietf.org/html.charters/diffserv-charter.html

[15] DPSS
The Distributed-Parallel Storage System (DPSS) is a scalable, high-performance, distributed-parallel
data storage system developed in the MAGIC Testbed. The DPSS is a collection of wide area distributed disk servers which operate in parallel to provide logical block level access to large data sets. Operated primarily as a network-based cache, the architecture supports cooperation among independently owned resources to provide fast, large-scale, on-demand storage to support data handling, simulation, and computation in a high-speed wide-area network-based internetworked environment.
See http://www-didc.lbl.gov/DPSS

Where We Are Today

[16] GASS
Globus Access to Secondary Storage. See "The Globus Project: A Status Report." I. Foster, C. Kesselman, Proc. IPPS/SPDP '98 Heterogeneous Computing Workshop, pg. 4-18, 1998. http://www.globus.org/documentation/papers.html

[17] Globus
The Globus project is developing the fundamental technology that is needed to build computational grids, execution environments that enable an application to integrate geographically-distributed instruments, displays, and computational and information resources. Such computations may link tens or hundreds of these resources.
See http://www.globus.org

[18] Grid
"The Grid: Blueprint for a New Computing Infrastructure," edited by Ian Foster and Carl Kesselman. Morgan Kaufmann, Pub. August 1998. ISBN 1-55860-475-8

[19] HENP
"High-Speed Distributed Data Handling for HENP: Architecture and Implementation," William E. Johnston, William Greiman, Brian Tierney, Craig Tull Information and Computing Sciences Division Douglas Olson, Nuclear Science Division, LBNL.
Available at http://www-itg.lbl.gov/STAR/

[20] MAGIC
"The MAGIC Gigabit Network." See: http://www.magic.net

[21] Nexus
The Nexus multithreaded runtime system is the communication component of Globus.
See http://www.globus.org/documentation/papers.html

[22] Nile
Nile is developing a distributed computing solution for the CLEO High Energy Physics experiment. The goal is to provide a self-managing, fault-tolerant, heterogeneous system of hundreds of commodity workstations, with access to a distributed database in excess of 100~TB. These resources are spread across the United States and Canada at 24 collaborating institutions. Nile will allow any resource to be accessed and used transparently by any member of the collaboration, from anywhere within the collaboration.
http://www.nile.utexas.edu/Nile

[23] NTON
"National Transparent Optical Network Consortium." See http://www.ntonc.org.
NTON currently consists of an eight wavelength, 2.5 gigabits/sec/channel, WDM optical fiber ring "around" the San Francisco Bay Area. In the future it will be extended from Seattle to San Diego.

[24] OOFS
"Objectivity/HPSS Interface (oofs)," http://info.in2p3.fr/CC/hepix98/storage/oofs/ .

Where We Are Today

[25] QoS
"Bandwidth Reservation: QoS as Middleware," William E. Johnston and Gary Hoo. To be presented at NASA NREN QoS Workshop. NASA Ames, August, 1998.

[26] SRB
See "Metadata Design for a Massive Data Analysis System," C. Baru; R. Frost; J. Lopez; R. Marciano; R. Moore; A. Rajasekar; M. Wan. In Proc. CASCON '96. http://www.sdsc.edu/MassDataAnal/

[27] RIO
Remote I/O (RIO) is the remote data access capability in Globus.
See http://www.globus.org/documentation/papers.html

[28] STAF
See http://iago.lbl.gov/STAF/staf.html

[29] WALDO
"The Wide Area Large Data Object Architecture: We are exploring the use of highly distributed computing and storage architectures to provide all aspects of collecting, storing, analyzing, and accessing large data-objects. These data-objects can be anywhere from tens of MBytes to tens of GBytes in size. They are typically the result of a single operational cycle of an instrument, such as: single large images from electron microscopes, video images from cardio-angiography, sets of related images from MRI procedures and images and numerical from a particle accelerator experiment. The source of such data objects, e.g. centralized health care facilities or large scientific instruments is often remote from the users of the data and from available large-scale storage and computation systems."
"Our Large Data-object Architecture utilizes a high-speed wide-area ATM network between the object sources and a mutli-level distributed storage system (DPSS). As the data is being stored, a cataloguing system (ImgLib) automatically creates and stores condensed versions of the data, textual metadata and pointers to the original data. The catalogue system provides a Web based graphical interface to the data. The user is able the view the low-resolution data with a standard internet connection and Web browser, or if high-resolution is required can use a high-speed connection and special application programs to view the high-resolution original data."
See http://www-itg.lbl.gov/WALDO